Primer on Medical Genomics
Part IV: Expression Proteomics

ANIMESH PARDANANI, MD, PHD; ERIC D. WIEBEN, MD; THOMAS C. SPELSBERG, PHD; AND AYALEW TEFFERI, MD

Proteomics, simply defined, is the study of proteomes. More completely, proteomics is defined as the study of all proteins, including their relative abundance, distribution, posttranslational modifications, functions, and interactions with other macromolecules, in a given cell or organism within a given environment and at a specific stage in the cell cycle. Proteins carry out the biological functions encoded by genes; hence, once the initial stage of genome sequencing and gene discovery is completed, a study of the proteome must be undertaken to address fundamental biological questions. The 3 broad areas are expression proteomics, which catalogues the relative abundance of proteins; cell-mapping or cellular proteomics, which delineates functional protein-protein interactions and organelle-specific protein distribution; and structural proteomics, which characterizes the 3-dimensional structure of proteins. With these approaches, proteins are studied on a global scale using a synergistic combination of powerful, high-throughput technologies, including 2-dimensional polyacrylamide gel electrophoresis, mass spectrometry, multidimensional liquid chromatography, and bioinformatics. Mass spectrometry, which provides highly accurate molecular mass measurements, has emerged as the analytical technology of choice for protein identification, characterization, and sequencing. This task has been made considerably easier with the availability of complete, nonredundant, and annotated genome sequence databases for many organisms. This article reviews the area of expression proteomics.

***Mayo Clin Proc*. 2002;77:1185-1196**

cDNA = complementary DNA; CID = collision-induced dissociation; 2-D = 2-dimensional; EMBL = European Molecular Biology Laboratory; ESI = electrospray ionization; EST = expressed sequence tag; ExPASy = Expert Protein Analysis System; HES = hypereosinophilic syndrome; IPG = immobilized pH gradient; LC = liquid chromatography; MALDI = matrix-assisted laser desorption ionization; mRNA = messenger RNA; MS = mass spectrometry; MS/MS = tandem mass spectrometry; NRDB = Non-Redundant Database; PAGE = polyacrylamide gel electrophoresis; PDB = Protein Data Bank; pI = isoelectric point; PMF = peptide mass fingerprinting; SDS-PAGE = sodium dodecyl sulfate PAGE; TOF = time of flight; TQ = triple quadrupole; TrEMBL = Translation of the EMBL Nucleotide Sequence Database

Proteomics, a term first coined in 1994, is a somewhat amorphous term, with an evolving definition. By one description, it is the study of all protein forms in a given organism or cell type within a given environment and at a specific stage of the cell cycle. Such a study comprises the quantification and characterization of all proteins, including the enumeration of all posttranslational modifications and study of all functional protein-protein interactions. Traditional questions in science, such as "Which proteins are present in cells?" "In what abundance?" and "How do they interact within both signaling pathways and complex cellular networks to mediate specific cellular functions?" are now being addressed by a synergistic combination of powerful, high-throughput technologies, including 2-dimensional (2-D) polyacrylamide gel electrophoresis (PAGE), liquid chromatography (LC), mass spectrometry (MS), and bioinformatics.

In general, 3 broad areas of study in the field of proteomics can be distinguished. The first is termed *expression* or *abundance proteomics*, which catalogues the relative abundance of specific proteins in a given tissue. This approach aims to compare patterns of protein expression in a given tissue under certain conditions, for example, health and disease or localized and advanced disease. Such a comparison would potentially yield markers of disease or its progression and eventually targets for therapeutic intervention.

The second area is termed *cell-mapping* or *cellular proteomics*, which aims to delineate protein-protein interactions to compile a framework of the complex networks that constitute intracellular signaling pathways. This aim will likely have a greater impact than the first because, given the paucity of human genes (relative to other organisms), it is evident that functional evolution of proteins in higher eukaryotes has resulted in large part from combinatorial diversification of intracellular regulatory networks

From the Division of Hematology and Internal Medicine (A.P., A.T.) and Department of Biochemistry and Molecular Biology (E.D.W., T.C.S.), Mayo Clinic, Rochester, Minn. Drs Wieben, Spelsberg, and Tefferi are members of the Mayo Clinic Genomics Education Steering Committee.

Individual reprints of this article are not available. The entire Primer on Medical Genomics will be available for purchase from the *Proceedings* Editorial Office at a later date.

mediated through novel protein-protein interactions.[1] The anticipated use of protein microarrays will be a powerful tool in the study of these diverse interactions.

The third area is termed *structural proteomics*, which strives to be able to predict the 3-dimensional structure of all proteins through a broad sampling of the universal protein structure space. The premise is that if high-resolution structures are obtained for a sufficiently large number of proteins, then essentially all proteins can be modeled on the basis of a solved structure of a significantly homologous protein.[2] High-resolution structures of proteins are of obvious biomedical relevance because they define the active site and functional domains of proteins and enhance our overall understanding of structure-function relationships, including enzyme catalysis, protein stability, and interactions within multimolecular complexes, thus allowing for rational drug design for the treatment of disease states.

This review focuses on the area of expression or abundance proteomics since such an approach is currently in widespread use for the comparison of global protein expression profiles between normal and tumoral or diseased tissue or before and after treatment with a drug. It also sets the stage for the introduction and description of several technologies that are an integral component of current proteomics research.

GENOMICS TO TRANSCRIPTOMICS TO PROTEOMICS

Genomics was shown to be feasible after a substantial lag phase since the initial sequencing of an entire bacteriophage in 1982.[3] The first use of the term *proteome* coincided with the publication of the first complete genomic DNA sequence of a self-replicating, nonparasitic organism (*Haemophilus influenzae* in 1995). For several years, DNA sequencing has progressed considerably. Rapid progress in determining the complete genome sequence of model organisms, including the yeast *Saccharomyces cerevisiae* (May 1996),[4] the nematode *Caenorhabtidis elegans* (December 1998),[5] the fruitfly *Drosophila melanogaster* (March 2000),[6] and the plant *Arabidopsis thaliana* (December 2000),[7] culminated in publication of a working draft of the human genome sequence in February 2001.[1,8] (A central Internet resource for genome projects representing many organisms is the Web site maintained by the National Center for Biotechnology Information: www.ncbi.nlm.nih.gov/Genomes.)

One of the first steps in genomic sequence analysis is identifying genes. Unfortunately, the quality of gene prediction from genome sequence data using current computational algorithms is limited. Despite more than 30 genomes having been sequenced to date, the successful prediction of genes remains a problem. The larger genome size of higher vertebrates makes the task of gene discovery even more difficult. Genome sequencing studies indicate that nearly all the increase in the average gene size in humans compared with other organisms (such as the fruitfly and worm) is due to the noncoding sequences, called *introns*, becoming much longer (coding sequences comprise an average of 5% of the length of a typical human gene).[1] The protein coding regions, termed *exons*, on the other hand, are roughly the same size (average length is 1340 base pairs in humans). The resulting decrease in signal (exon)-to-noise (intron) ratio in the human genome increases the probability of misprediction of genes when computational ab initio gene-finding algorithms are used. The public International Human Genome Sequencing Consortium has been conservative in that it has required that exons of genes predicted by computational strategies be confirmed by demonstrating significant homology to previously known genes or protein sequences.[1] In this regard, analysis of DNA sequences that are expressed in cells (represented in complementary DNA [cDNA] or expressed sequence tag [EST] libraries) and proteins from humans and other organisms provide a powerful resource for the prediction of genes. Such approaches, however, are not helpful in the identification and discovery of thousands of human genes that produce non-(protein) coding RNAs (such as transfer RNAs, ribosomal RNAs, small nuclear and nucleolar RNAs). Spurious prediction may also pose a problem with such an approach. It appears that nongene DNA sequences are transcribed frequently.[9] These transcripts either cannot be translated into a functioning protein or encode for proteins that are nonfunctional and consequently degraded rapidly.

A surprising finding has been the relative paucity of genes in the human genome, probably between 30,000 and 60,000, only about twice the number found in other organisms, such as the fruitfly *D melanogaster*, the worm *C elegans*, or the plant *A thaliana*.[1,8] The phenomenon of alternative splicing of messenger RNA (mRNA) appears to be more prevalent in humans than in other species.[1] In other words, there are often many more ways in which a gene's protein coding regions (exons) can be joined to create a functional mRNA molecule, ready to be translated into protein. Alignment of ESTs to the working draft sequence suggests that 60% of human genes have multiple splicing variants.[1]

Also, there are numerous ways in which eukaryotic proteins can be posttranslationally modified. This implies that more protein isoforms are encoded per gene in humans than in other species. In other words, more proteins comprise a proteome than genes a genome. This constitutes a substantial departure from the central tenet of "one gene, one protein." Individual proteins are composed of discrete structural units called *domains*, which have critical functional or structural roles and are thus stringently conserved during evolution. Although most domains and motifs in human proteins are also present in proteins from other organisms (only an estimated 7% are vertebrate specific), they have been shuffled to create many more different combinatorial arrangements (an estimated 2-5 times as many) in humans.[1,8] This combinatorial diversification is most prominent in the evolution of novel extracellular and transmembrane protein architectures.[1]

Thus, vertebrate evolution from invertebrates appears to have required the invention of few new domains. These observations imply that comparative genomics is likely to prove a powerful tool in the gene prediction toolbox.[10] The comparison of conserved sequences between human and other vertebrate genomes (or between other closely related genomes) will allow for the detection of novel genes, even when their function is unknown in either species. Sequencing of the mouse, zebrafish, pufferfish, and other genomes holds great promise for the utility of comparative genomics in gene prediction.

Within all cells, DNA is a relatively stable, heritable informational molecule. In fact, the molecular machinery of cells is geared toward maintaining the integrity of DNA in the face of various intrinsic and extrinsic insults. Hence, the mere presence of a gene within a cell reveals little about its functional competence. In other words, whether and when a particular gene is transcribed and translated, at what rate, under which specific circumstance, and its functional consequence(s) remain undetermined from studies at the level of genomic sequence. Additionally, many genes are pseudogenes that are no longer expressed in cells. Once the initial stage of genome sequencing and gene discovery is completed, attention must be focused on gene expression and the functions of the proteins they encode. Such questions, which lie at the heart of what is now termed *functional genomics*, must be answered downstream to the level of DNA sequence (ie, at the level of gene-encoded mRNA transcripts and proteins).

The study of gene expression can be undertaken at the level of mRNA transcripts. Transcriptomics, the study of the complete set of cellular transcripts under defined conditions, represents the second major area of genome science analysis. (Specific instruments and techniques have been detailed previously.[11,12]) Technologies such as cDNA and oligonucleotide microarrays exist that facilitate the parallel, quantitative analysis of the expression of thousands of genes.

Such an experimental approach, however, also has inherent limitations, which include the following:

1. Sensitivity: Weakly expressed, low–copy number mRNA transcripts may not be detected. This has implications for the choice of array platform (eg, fluorescent or radioactivity based).

2. Specificity: Discrimination between alternatively spliced or closely related (highly homologous) transcripts may be difficult (particularly with cDNA microarrays).

3. Quantitative: It may be difficult to achieve equally high stringency of hybridization across the entire high-density array.

These limitations demand a post hoc confirmation of experimental results. Large-scale microarray experiments frequently serve as an initial "hypothesis generation" step. Given the statistical challenges in distinguishing low-magnitude (less than 2-fold) changes in gene expression from random chance, corroborating experiments (frequently at the protein level, eg, immunoblotting or other proteomic approaches) are usually necessary to establish the biological importance of these observed changes.

The fundamental issue, however, is whether the study of transcriptomes represents the ultimate complement of genomics for the analysis of gene function. If the abundance and activity of the end product of genes (ie, proteins) are determined exclusively or primarily by regulatory events at the level of gene expression, then transcriptomics would represent an optimal approach for the study of functional genomics. However, this clearly is not the case, and it is now generally agreed that analyses undertaken at the level of proteins are necessary for the following reasons.

First, there is poor correlation between mRNA abundance and corresponding protein levels[13,14] (correlation factor of about 0.4). This implies that protein levels cannot simply be predicted from corresponding mRNA levels.

Second, virtually all eukaryotic proteins undergo posttranslational modification(s). These modifications, which potentially have enormous functional consequences, cannot always be predicted from gene sequences.

Third, the translocation of protein from its site of synthesis to the site of activity cannot always be deduced from sequence data.

Fourth, protein function cannot always be reliably predicted from sequence information.

The term *proteome* was first proposed in 1994 at the Siena 2-D Electrophoresis meeting to depict the *prot*ein complement of a gen*ome*. In the complex interplay of molecular events that lead from gene activation to the synthesis of functionally competent protein, the proteome represents the end product of the genome. Although the cellular genome is relatively constant, the proteome is in constant flux. The global protein complement of a given cell varies with changes in the physiological state of the cell and its ambient environment. These changes include activation of specific cellular signaling pathways, position in the cell cycle, and drug exposure. Moreover, different

Table 1. **Summary of the Advantages and Disadvantages of Select Methods***

Method	Advantages	Disadvantages
2-D PAGE	Theoretically may resolve up to 10,000 cellular proteins; considerable commercial support for 2-D PAGE systems is available	Labor and time intensive; inherent run-to-run variability; limited dynamic range; biased toward display of cytosolic and hydrophilic proteins
MALDI-TOF MS (for PMF)	Spectra are relatively simple to interpret; large mass range; high sensitivity; low sensitivity to salts and other contaminants	Success of PMF depends on access to complete, nonredundant, and annotated DNA sequence databases; PMF may fail because of unanticipated peptide posttranslational or artifactual modifications, or nonspecific proteolysis
MS/MS (for peptide (fragmentation)	Derived peptide sequence tags facilitate accurate and large-scale protein identification	Sequence-specific peptide fragmentation spectra may be difficult to interpret; success depends on access to complete, nonredundant, and annotated DNA sequence databases
LC-MS/MS	Complex peptide mixtures can be analyzed; high-throughput analysis is possible	As listed for MS/MS

*2-D PAGE = 2-dimensional polyacrylamide gel electrophoresis; LC = liquid chromatography; MALDI-TOF = matrix-assisted laser desorption ionization time of flight; MS = mass spectrometry; MS/MS = tandem MS; PMF = peptide mass fingerprinting.

cell types within a multicellular organism will have different proteomes. Although the proteome of a given cell at any moment represents only the expression of part of the genome, proteomes are complex.

Although the basic building blocks of DNA (the 4 nucleotides) are relatively homogeneous in terms of their chemical composition, the comparable protein components are far more complex (22 unmodified[15] and many more modified amino acids for proteins). Similarly, DNA structure is relatively homogeneous compared with the incredibly broad diversity of protein structures. Additionally, genes can be variably spliced, mRNA editing occurs commonly, and almost all eukaryotic proteins are cotranslationally and/or posttranslationally modified in a variety of ways. This complexity explains why proteomics must be considered a science as large-scale as genomics, if not even more so. An additional level of complexity in proteome research exists because of the absence of technologies, such as polymerase chain reaction and cloning, that facilitate the amplification and separation of biological macromolecules of interest. These are crucial components of any large-scale analytic process and are not easily achieved for protein analysis.

Thus, protein analysis and defining a cell's proteome are challenging endeavors. It is intuitively obvious that a truly holistic approach toward analysis of basic biological questions will use a method that studies molecular events comprehensively at the level of the genome, transcriptome, and proteome.

PROTEOME ANALYSIS

Ideally, proteome analysis (proteomics) should characterize and quantify all proteins in a specific cell type under a specific set of environmental conditions, including all posttranslational modifications. Currently, the combination of 2-D PAGE and MS comes closest to realizing this goal in the real world, although complementary methods based on LC have been developed. Protein microarrays similar to gene arrays promise a new approach (and tool) in proteome analysis. Such a large-scale analysis involves the steps of initial protein separation and subsequent protein characterization. The pros and cons of select methods are summarized in Table 1.

Protein Separation

Since its initial description more than 25 years ago,[16-18] protein separation has traditionally been achieved using 2-D PAGE, wherein separation according to charge (isoelectric point [pI]) is followed by separation according to molecular mass (Figure 1). Several important advances, including the introduction of immobilized pH gradients (IPGs) in the 1980s,[20,21] which produce stable and reproducible pH gradients for first-dimension separation, have made it possible to resolve complex protein mixtures in a reproducible manner. Additional advances include the following.

The first advance is the development of newer cocktails that enhance solubilization of membrane and other hydrophobic proteins,[22] which were previously poorly resolved (only approximately 1% of integral membrane proteins were conventionally resolved).[23]

The second advance is an increased resolution of proteins in complex mixtures with the use of narrow pI-range ("zoom") gels that cover narrow pH ranges (usually 1 pH unit).[24,25] This facilitates visualization of a smaller but more detailed window of the proteome. With the use of overlapping narrow-gradient gels, it is possible to resolve more than 10,000 proteins from a higher eukaryotic cell lysate.[26]

The third advance is the development of fluorescent protein dyes (eg, Sypro ruby, red, and orange)[27] that have comparable sensitivity (nanogram range) to silver stain[28] (probably the most popular and sensitive nonradioactive protein detection stain) and a larger dynamic range (>10^4) in protein visualization. These fluorescent dyes are also compatible with downstream MS analyses.

The fourth advance is the availability of proprietary 2-D image analysis software that allows large-scale comparisons of differential protein expression by tracking the numerous protein spots in sets of comparative 2-D gels.

The fifth advance is the automation and high-throughput analysis of protein expression.[29] Although the commercial support for 2-D PAGE systems is substantial and equipment is available to run several large-format gels in parallel, this approach is cumbersome, and 2-D image analysis, spot excision, protein digestion, and protein extraction can hinder the throughput of the whole process.

Two-dimensional PAGE has several limitations.[30,31] It produces a snapshot of a cell, the quality and biological relevance of which reflects the upfront experimental design and sample preparation. For example, if 2-D PAGE maps from primary and metastatic tumor tissue are compared to identify proteins whose abundance varies with tumor progression, it is necessary to consider differences arising from patient-to-patient heterogeneity and the multiplicity of cell types that exist within the tumor. Sample preparation is a critical step in 2-D PAGE, and in attempting to compare relative protein abundance in 2 samples, standardization of sample preparation protocols and loading of comparable amounts of sample across comparison sets are important considerations.

In addition, 2-D gels exhibit inherent run-to-run variability in the observed pattern of protein spots. This complicates the comparative analyses of 2-D gels, particularly between laboratories and with archived images in 2-D databases.

Moreover, 2-D analysis has a limited dynamic range, and low–copy number proteins are visualized rarely. It is estimated that the dynamic range of silver-stained gels is 3 orders of magnitude. This is seriously limiting, considering the actual dynamic range of intracellular proteins is about 7 to 8 orders of magnitude.[32] Thus, low-abundance proteins tend to be poorly represented in 2-D gels. Increasing the protein load in an attempt to visualize and identify low-abundance proteins has not proved to be a successful strategy because the load capacity of the gels (IPG strips) is frequently exceeded, and consequently a poor resolution of proteins is obtained.

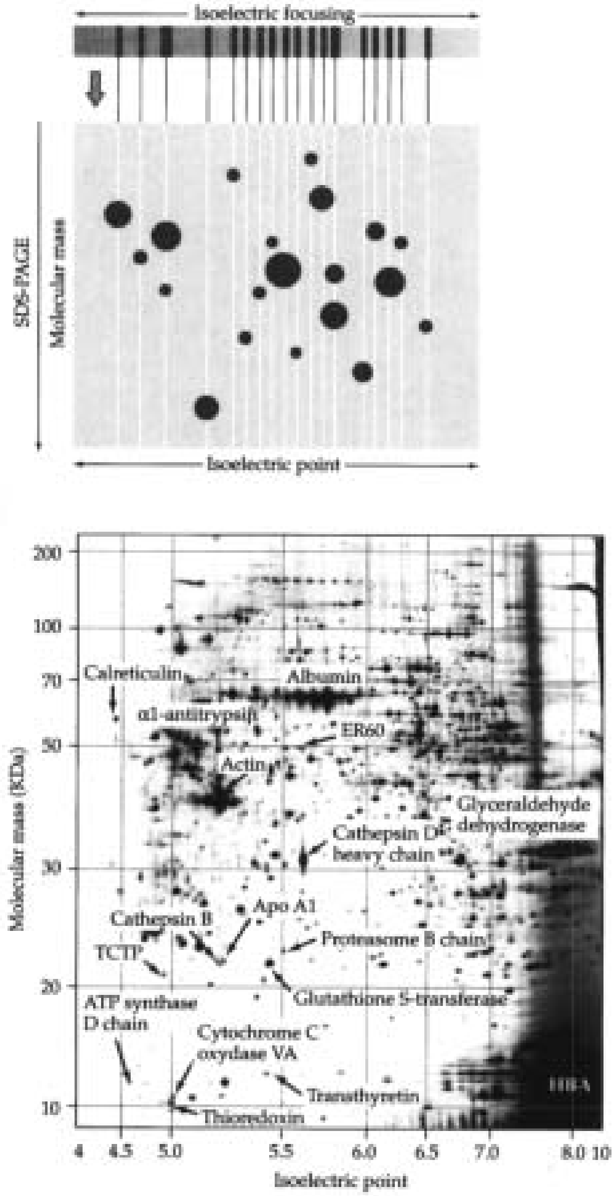

Figure 1. Two-dimensional polyacrylamide gel electrophoresis (2-D PAGE). Top, In 2-D PAGE, protein extracts are applied to the center of an isoelectric focusing strip and allowed to diffuse along the ionic gradient to equilibrium. The strip is then applied to a sodium dodecyl sulfate (SDS) gel, in which electrophoresis in the second dimension separates proteins into “spots” according to molecular mass. The size of each spot is proportional to the amount of protein. Bottom, A partially annotated human lymphoma 2-D gel from the ExPASy (Expert Protein Analysis System; www.expasy.org/) database. Reprinted with permission from Gibson and Muse.[19]

Finally, most 2-D gel analyses of cell extracts are selectively biased toward primarily cytosolic and hydrophilic proteins. This results from nonstoichiometric extraction of proteins from cells. Hydrophobic or membrane proteins, low-abundance proteins, proteins whose pI values are at the extremities of pH gradients (very acidic or basic proteins), and proteins sequestered in organelles tend to be

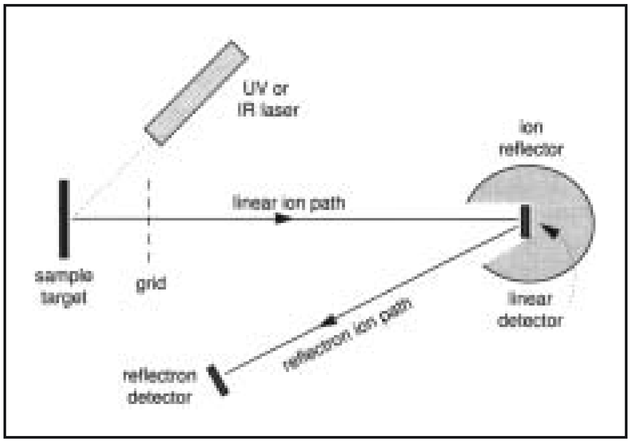

Figure 2. Schematic diagram of a matrix-assisted laser desorption ionization (MALDI) time-of-flight (TOF) mass spectrometer. Microliter quantities of liquid samples are mixed with a matrix molecule and dried onto a stainless steel or gold-plated target. A pulsing laser is used to irradiate the matrix-embedded sample. This creates molecular ions that are accelerated by an electric field that exists between the sample target and grid. Ions then enter a field-free flight tube with a velocity essentially proportional to their mass, and their TOF in this tube is measured at the linear detector. Small ions fly faster than larger ions, and their mass-to-charge (m/z) ratios can be calculated from their flight time by using compounds of known mass as calibrants. To increase mass resolution, some MALDI instruments have an ion reflector. This turns ions around in an electric field, sending them toward a second detector. IR = infrared. Reprinted with permission from Wilkins and Gooley.[48]

poorly represented in 2-D PAGE gels relative to their abundance within cells.

Protein Identification and Characterization: Correlating MS Data With Sequence Databases

A traditional approach toward characterization of protein structure and function has been to clone the gene encoding for the protein. This is accomplished by determination of partial protein sequence by de novo N-terminal sequencing (Edman degradation),[33] which facilitates the synthesis of degenerate oligonucleotide primers as a means to screen appropriate cDNA libraries. However, for those species for which the complete genome has been sequenced, protein identification is potentially more straightforward because essentially all possible protein sequences are represented in the genomic sequence database. Another source of extensive nucleotide sequence information is the result of EST sequencing efforts. The ESTs are short sequences of 250 to 400 base pairs generated by random single-pass sequencing of cDNA libraries. Given the access to extensive nucleotide sequence information currently available, all potential protein sequences in a given organism can be deduced from 6-way translations of the nucleotide sequence. Protein identification then, in large part, involves the matching of experimentally derived protein attributes, such as pI, molecular mass, and amino acid sequence, against those predicted from the translation of genomic or cDNA sequences in databases. Software programs generate a list of protein entries, ordered by a score that reflects the fit between theoretical and experimental parameters.

An analytical technique that can provide highly accurate molecular mass measurements of molecules or their component fragments, MS is now firmly established as the method of choice for protein identification and characterization. (Specific instruments and techniques have been discussed previously.[34-43]) In addition, MS is continuing to evolve rapidly in terms of its sensitivity (femtomole), accuracy (sub-10 ppm), and diversification into an array of technologies adapted to specific applications in protein identification and characterization.

A mass spectrophotometer consists of 3 components: (1) an ionization source, (2) a mass analyzer, and (3) a detector. Mass spectral analysis requires that the analyte be introduced into the mass spectrometer as a gaseous ion. Two "soft" ionization methods that convert large biological polymers (>10,000 kd) into gaseous ions were introduced a decade ago. The first, matrix-assisted laser desorption ionization MS (MALDI-MS),[44,45] generates ions from solid-phase samples, whereas the second, electrospray ionization MS (ESI-MS),[46,47] generates ions from liquid-phase samples, usually following elution off a chromatography column that separates complex protein mixtures. Commonly used mass analyzers are a flight tube or time of flight (TOF) and a quadrupole. Although various combinations of ionization sources and mass analyzers are possible, the most common are MALDI-TOF MS (Figure 2), ESI-TQ (triple quadrupole) MS (Figure 3), and ESI-TOF MS.

Two types of MS data have been used for protein identification by matching experimental attributes with those predicted from sequence databases: (1) the accurate mass of peptides (within 5-ppm resolution) derived by the sequence-specific proteolysis of the target protein (peptide mass fingerprinting [PMF]) and (2) MS fragmentation spectra (primarily y and b ions) allowing for de novo peptide sequencing.

The technique of PMF, currently the most common method used to identify proteins in a high-throughput environment, was described independently by several groups in 1993.[49-52] It can be summarized (Figure 4) as follows.

As a first step, the protein(s) of interest is isolated (by 2-D PAGE) and cleaved either in gel or on membranes by specific enzymatic or chemical methods. A commonly used enzyme is trypsin that cleaves only at the C-terminal side of arginine or lysine. Next, masses of resulting peptides, which correspond to the specific amino acid sequence of an individual protein, are measured at high resolution and mass accuracy in a mass spectrometer. Although the spectrophotometric technique of choice for measuring peptide masses is debatable, MALDI-TOF MS is emerging as the best alternative for rapid screening. Finally, a nonredundant protein or translated nucleotide sequence database is screened, and a list of theoretical peptide masses is generated for every protein by cleaving it in the manner of the experimentally used enzyme (eg, trypsin). A ranking or score is then calculated to provide a measure of the fit between the observed and predicted peptide masses. The correct identification of an unknown protein is likely to be that candidate with the largest number of peptide "hits." Confidence in protein identification is derived from a high spread in scores when the top-ranked protein is compared with secondary matches.

Although PMF is a powerful tool for rapid and sensitive protein identification, analysis of the results of PMF experiments may not always be straightforward. In many cases there are orphan peptide masses that do not match up with peptide masses predicted from the highest-ranked candidate protein(s). The success of PMF experiments is maximized when peptide masses are computed by using full-length protein or gene databases (vs short-sequence EST databases) and when high peptide mass accuracy is available. The probability of accurate protein identification is highest when additional attributes, other than experimental peptide masses and the proteinase used to digest the proteins, are specified to constrain the search algorithm. These optional attributes may include the following information: species of origin, molecular weight and/or pI of the protein, known posttranslational (eg, phosphorylation) or potential artifactual modifications (such as oxidation of methionines), protein N- and C-terminal sequence tags, and amino acid composition. The minimum number of matching peptides required for a protein to be suggested as a possible match may also be specified as a parameter. Additional methods that increase the information content of individual peptides (orthogonal approaches that include chemical modification of peptides, such as esterification and hydrogen-deuterium exchange) can also be used. When sample information, which is made available to the search algorithm, is maximized, the search space is reduced proportionally, thereby decreasing the number of candidate proteins and decreasing the probability of false matches. However, one must be careful not to miss the correct protein by overly constraining the search space.

Most proteins from higher eukaryotes undergo cotranslational and/or posttranslational modifications. These include the addition or removal of simple chemical groups (eg, hydroxyl-, phosphoryl-, carboxyl-) or a cleavage process (eg, signal sequences, propeptides, initiator methionines) or the addition of more complex moieties, such as sugars and lipids. In these situations, peptide masses that are computationally predicted by using nonannotated genome sequence databases often fail to match the experimentally determined peptide masses. Furthermore, such a matching of peptide masses may fail because of nonspecific proteolysis of the sample, cleavage by a contaminating protease, or contamination of the protein under study by other proteins. Such errors that may arise during sample processing also need to be considered. Rarely, peptide matching may fail if a truly novel protein has been isolated. Fortunately, powerful and comprehensive software tools for the discovery of protein posttranslational modifications and identification of possible peptides that have resulted from nonspecific chemical or enzymatic cleavage of proteins are now available online (http://us.expasy.org/tools/#ptm).

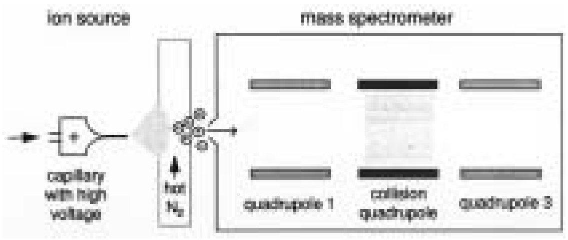

Figure 3. Schematic diagram of electrospray ionization mass spectrometer, in this case a triple quadrupole mass spectrometer equipped with an electrospray ion source. Peptide ions are introduced initially via the electrospray source into the first quadrupole, where peptide masses can be measured. If desired, chosen ions can then be sent to the second quadrupole for fragmentation by collision with an inert gas (nitrogen [N_2]). The mass of fragmentation products can then be measured in the third quadrupole. Reprinted with permission from Wilkins and Gooley.[48]

MALDI-TOF MS is emerging as the best alternative for rapid screening of peptides for PMF.[53] There are many reasons for this.

First, MALDI-TOF MS produces singly charged ions (ie, each peptide carries only 1 charge and hence generates only 1 peak in the spectrum). This makes MALDI-TOF MS data relatively easy to interpret.

Second, MALDI-TOF MS has a large mass range (500-600 d up to a few hundred thousand daltons). Thus, it is capable of analyzing whole proteins and other polymers.

Third, MALDI-TOF MS has a short analysis time, high sensitivity (50-100 fmol range, such that a fraction of a Coomassie blue–stained spot from a gel is sufficient mate-

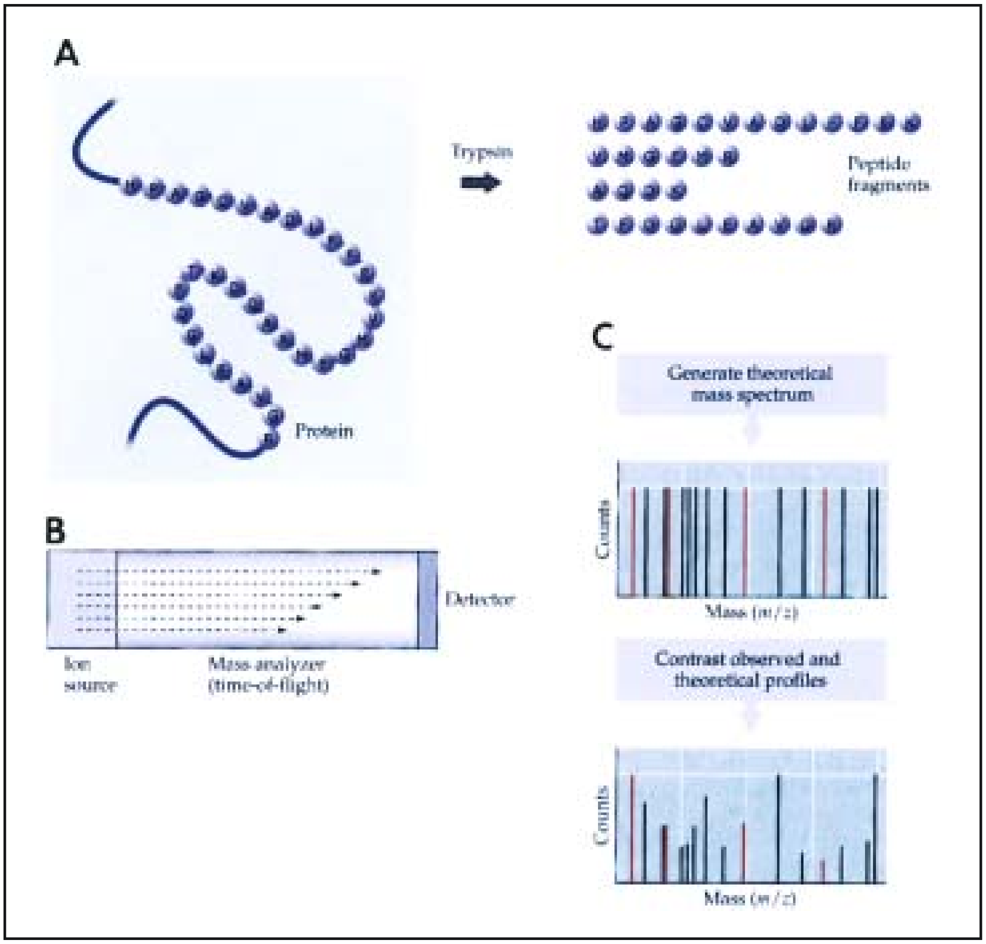
A
Trypsin
Peptide fragments
Protein
B
Detector
Ion source
Mass analyzer (time-of-flight)
C
Generate theoretical mass spectrum
Counts
Mass (m/z)
Contrast observed and theoretical profiles
Counts
Mass (m/z)

Figure 4. Peptide mass fingerprinting. A, Proteins isolated from a 2-dimensional polyacrylamide gel electrophoresis gel or chromatography column are digested by trypsin into short peptide fragments. B, Fragments are separated by matrix-assisted laser desorption ionization time-of-flight mass spectrometry. C, Matching of profiles of multiple peptides in the observed mass spectrum with all the predicted peptide fragments for every predicted protein allows protein identification. m/z = mass-to-charge ratio. Reprinted with permission from Gibson and Muse.[19]

rial for analysis), and high mass accuracy (<30 ppm) as well as high resolution of peptides, which allows stringent mass constraints to be specified for PMF.[54]

Finally, MALDI-TOF MS displays low sensitivity to salts and other contaminants, which allows peptides taken from in-gel or on-membrane digests to be analyzed directly.

Several software tools are available to identify proteins using PMF. These programs generate a list of protein entries, ordered by a scoring system that reflects the fit between theoretical and experimental parameters.

Peptide mass value is the only attribute by which PMF characterizes each peptide. By itself, however, a mass value has little informational content about the peptide sequence. Protein identification by PMF, which strongly depends on access to complete and nonredundant (DNA, EST, or protein) databases, may fail to identify candidate proteins because of the limited availability of such data for some species. Additional reasons for failure include searches against extremely large sequence databases or if peptides have posttranslational or artifactual modifications.[55,56] For such samples, additional information is necessary for unambiguous protein identification.

Fragmentation of peptide ions by MS provides extensive amino acid sequence–specific information of the protein to enable identification by searching against protein sequence databases. One approach is tandem MS (MS/MS), in which peptide fragmentation (Figure 5, left) is achieved by collision-induced dissociation (CID) accomplished by ESI-TQ MS (Figure 3) or by ion-trap MS/MS. The informational content of a CID spectrum in an MS/MS experiment is essentially a plot of the frequency of peptide fragmentation (residue mass weights of a parent peptide ion in the form of smaller ion fragments, primarily y and b ions). These spectra can provide rapid and unambiguous identification of a peptide from sequence databases because of the complementary and redundant nature of data present in such spectra.

Automated computational algorithms similar to those for PMF can be used to search protein or translated nucleotide sequence databases with uninterpreted CID MS/MS spectra that are derived experimentally.[57,58] Initially, all proteins contained in a database are theoretically digested to find parent peaks. Next, the theoretical parent peptides are computationally fragmented (into y and b ion spectra), and the experimental, uninterpreted CID MS/MS spectra are compared with the theoretical spectra. Similar to PMF, a scoring system is then used to reflect the degree of fit between theoretical and experimental spectra. The results of comparison of MS/MS fragmentation spectra can be improved by combining these data with peptide sequence tags. The identification of proteins with MS/MS is a powerful technique, especially for protein mixtures.[59]

Frequently (for a variety of reasons previously discussed) a protein will remain unidentified through database matching strategies even when high-quality, experimentally derived MS/MS spectra are available. In such a situation, an alternative is de novo interpretation of experimental MS/MS spectra; ionic fragments resulting from peptide fragmentation produce a ladder of peptides, in which the mass difference between ionic fragments corresponds to specific amino acids (Figure 5, right). Peptide fragmentation, which can be controlled precisely by CID in an ESI-TQ MS instrument (Figure 3), generates a series of ions from which sequence can be deduced. A few residues of sequence obtained from a CID spectrum, called a *peptide sequence tag*, combined with peptide parent ion mass may be sufficient to identify a protein.[58,60] However, the interpretation of MS/MS spectra is not always simple, and it is rare that complete peptide sequences can be deduced de novo from such spectra. In many cases, data from a single MS/MS spectrum may be insufficient to determine unambiguously the peptide sequence, and additional methods, such as fragmentation of specifically derivatized peptides, may assist in de novo sequencing in such situations.

DATABASES RELEVANT TO PROTEOME RESEARCH

A vast number of general or specialized proteome databases are available to researchers throughout the world.[61-63] Current computational and network technologies allow this vast amount of biological data to be integrated and extracted efficiently. This section describes a few selected databases relevant to proteome research to provide a glimpse into the kind of information available at such sources. Database descriptions are abstracted from the appropriate Web sites in most cases, for example, SWISS-PROT (www.expasy.org/),[64] TrEMBL (Translation of the EMBL [European Molecular Biology Laboratory] Nucleotide Sequence Database) (www.expasy.org/),[64] NRDB (Non-Redundant Database) (www.ncbi.nlm.nih.gov/), GenBank (http://ncbi.nlm.nih.gov/Genbank/), PROSITE (http://us.expasy.org/prosite/), 2-D PAGE databases, and 3-dimensional structure databases.

SWISS-PROT

The first molecular biology World Wide Web server, called ExPASy (Expert Protein Analysis System), allows queries to the SWISS-PROT database, which is the major database for curated protein sequences. This database distinguishes itself from other protein sequence databases by 3 separate criteria.

The first criterion is annotation. For each sequence entry, the annotation describes many features, including the function(s) of the protein, posttranslational modification(s), domains and sites, secondary and quaternary structure, similarities to other proteins, disease(s) associated with deficiencies in the protein, sequence conflicts, and variants.

The second criterion is minimal redundancy. Many sequence databases contain, for a given protein sequence, separate entries that correspond to different literature reports. In SWISS-PROT, all these data are merged to minimize the redundancy of the database.

The third criterion is integration with other databases. SWISS-PROT is currently cross-referenced with approximately 60 different databases. Currently, SWISS-PROT (release 40.12, March 5, 2002) has 105,967 entries.

TrEMBL

The TrEMBL database, available through the ExPASy server, supplements SWISS-PROT and can be considered a preliminary repository in which protein sequences derived by the translation of coding regions in the EMBL Nucleotide Sequence Database are stored before being manually annotated and moved to the main SWISS-PROT database.

NRDB

The NRDB is maintained by the National Center for Biotechnology Information. Strictly speaking, its entries are nonidentical but highly redundant, given the relatively primitive exclusion criteria for its entries. For example, multiple entries exist for a single protein when these entries differ due to sequencing errors and/or polymorphisms. These features limit the usefulness of NRDB and similar databases, such as OWL.

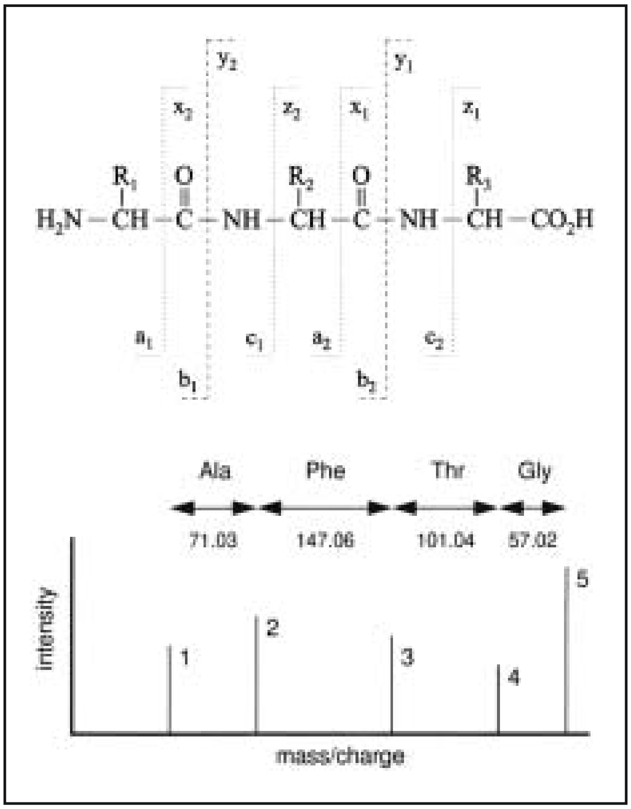

Figure 5. Top, Peptide fragmentation by tandem mass spectrometry. This figure shows how a peptide of 3 amino acids can fragment into a series of daughter ions. Fragmentation at the peptide bonds will produce b series ions if the peptide charge remains at the peptide N-terminus or y series ions if the charge remains at the C-terminus. Only the charged portion of the peptide will be detected after fragmentation. Other peptide fragmentations (not shown) can also occur. Bottom, De novo peptide sequencing. Peptide fragmentation is most useful when the ionic fragments produce a ladder of peptides, in which the mass difference between each peptide (arrows and values) corresponds exactly to that of a certain amino acid. In this manner, partial sequence of a peptide can be read. In this schematic diagram, a sequence of AFTG (alanine, phenylalanine, threonine, glycine) is present. Peptide 5 would have been the peptide XXAFTG, peptide 4 XXAFT, peptide 3 XXAF, peptide 2 XXA, and peptide 1 XX. Ala = amino acid alanine; Gly = glycine; Phe = phenylalanine; Thr = threonine. Reprinted with permission from Wilkins and Gooley.[48]

GenBank

GenBank is the National Institutes of Health genetic sequence database, an annotated collection of all publicly available DNA sequences. Sequence entries to this database are the responsibility of submitting researchers, and since this is a noncurated database, it is highly redundant. GenBank continues to grow at an exponential rate, with the number of nucleotide bases doubling approximately every 14 months. Currently, GenBank contains more than 17 billion bases from more than 100,000 species.

PROSITE

PROSITE is a database of protein families and domains. It is based on the observation that, even though there is a huge number of different proteins, most of them can be grouped on the basis of similarities in their sequences into a limited number of families. By analyzing the constant and variable properties of such groups of similar sequences, it is possible to derive a signature for a protein family or domain that distinguishes its members from all other unrelated proteins. A protein signature can be used to assign a newly sequenced protein to a specific family of proteins and thus to formulate hypotheses about its function. PROSITE currently contains patterns and profiles specific for more than a thousand protein families or domains.

Other related databases that use alternative pattern recognition algorithms to group proteins into domains or families, and hence infer function, include PFAM, BLOCKS, ProDom, PRINTS, and InterPro.

The 2-D PAGE Databases

SWISS-2DPAGE (http://us.expasy.org/ch2d/) is an annotated 2-D PAGE and sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS-PAGE) database. The SWISS-2DPAGE database assembles data on proteins identified on various 2-D PAGE and SDS-PAGE maps. In addition to textual data, SWISS-2DPAGE provides several 2-D PAGE and SDS-PAGE images that show the experimentally determined location of the protein. Cross-references are provided to MEDLINE and other federated 2-D databases (eg, COMPLUYEAST-2DPAGE, ECO2DBASE, and Siena-2DPAGE) and to SWISS-PROT, which provides many links to other molecular databases.

The 3-Dimensional Structure Databases

The Research Collaboration for Structural Bioinformatics Consortium maintains the Protein Data Bank (PDB) (www.rcsb.org/pdb/) database, the single worldwide archive of structural data of biological macromolecules, including proteins and nucleic acids. At last available update (April 23, 2002), there were 17,902 structures deposited in PDB. A variety of information associated with each structure is available, including sequence details, atomic coordinates, crystallization conditions, 3-dimensional structure neighbors computed using various methods, derived geometric data, structure factors, 3-dimensional images, and a variety of links to other resources.

CLINICAL EXAMPLE

We recently reported the efficacy of imatinib mesylate (Gleevec in the United States [Novartis]), an orally bioavailable inhibitor of specific tyrosine-kinases, for the hypereosinophilic syndrome (HES).[65] In a subsequent pilot study, we are investigating drug efficacy and mechanism of action of imatinib in a spectrum of hematologic disorders, including HES, that share the common characteristic of peripheral blood eosinophilia. Because the molecular pathogenesis of HES is currently unknown and imatinib is efficacious at a relatively low dose (100 mg/d), we suspect that imatinib's activity may be due to inhibition of a yet unidentified cellular target in these patients. To clarify this issue, we are analyzing the phosphoprotein complement of clonal cells in a comparative fashion. We are prospectively collecting pretreatment and posttreatment (at baseline, 24 hours, and 1 month) peripheral blood samples from which eosinophils are purified. Proteins phosphorylated at tyrosine residues are enriched from eosinophil cell lysates by immunoprecipitation using a cocktail of high-affinity anti-phosphotyrosine antibodies. The enriched proteins will be resolved by 1-dimensional and/or 2-D PAGE and pretreatment and posttreatment gels compared to identify differentially phosphorylated spots or bands. These spots or bands will be isolated and subjected to PMF in a preparative scale experiment. Peptides that fail to be identified by PMF will be subjected to fragmentation by MS/MS for sequencing. We hope that such an approach will shed light on the signaling pathways that are modulated by imatinib therapy.

REFERENCES

1. Lander ES, Linton LM, Birren B, et al, International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome [published corrections appear in *Nature*. 2001;412:565 and 2001;411:720]. *Nature*. 2001;409:860-921.
2. Chance MR, Bresnick AR, Burley SK, et al. Structural genomics: a pipeline for providing structures for the biologist. *Protein Sci*. 2002;11:723-738.
3. Sanger F, Coulson AR, Hong GF, Hill DF, Petersen GB. Nucleotide sequence of bacteriophage lambda DNA. *J Mol Biol*. 1982; 162:729-773.
4. Goffeau A, Barrell BG, Bussey H, et al. Life with 6000 genes. *Science*. 1996;274:546, 563-567.
5. *C. elegans* Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology [published corrections appear in *Science*. 1999;283:35, 2103 and 1999;285:1493]. *Science*. 1998;282:2012-2018.
6. Adams MD, Celniker SE, Holt RA, et al. The genome sequence of *Drosophila melanogaster*. *Science*. 2000;287:2185-2195.
7. Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*. 2000;408:796-815.
8. Venter JC, Adams MD, Myers EW, et al. The sequence of the human genome [published correction appears in *Science*. 2001; 292:1838]. *Science*. 2001;291:1304-1351.
9. Normile D, Pennisi E. Team wrapping up sequence of first human chromosome. *Science*. 1999;285:2038-2039.
10. Rubin GM. The draft sequences: comparing species. *Nature*. 2001; 409:820-821.
11. Grigorenko EV, ed. *DNA Arrays: Technologies and Experimental Strategies*. Boca Raton, Fla: CRC Press; 2002.
12. Jordan BR, ed. *DNA Microarrays: Gene Expression Applications*. Berlin, Germany: Springer-Verlag; 2001.
13. Anderson L, Seilhamer J. A comparison of selected mRNA and protein abundances in human liver. *Electrophoresis*. 1997;18:533-537.
14. Gygi SP, Rochon Y, Franza BR, Aebersold R. Correlation between protein and mRNA abundance in yeast. *Mol Cell Biol*. 1999;19: 1720-1730.
15. Atkins JF, Gesteland R. Biochemistry: the 22nd amino acid. *Science*. 2002;296:1409-1410.
16. Klose J. Protein mapping by combined isoelectric focusing and electrophoresis of mouse tissues: a novel approach to testing for induced point mutations in mammals. *Humangenetik*. 1975;26: 231-243.
17. O'Farrell PH. High resolution two-dimensional electrophoresis of proteins. *J Biol Chem*. 1975;250:4007-4021.
18. Scheele GA. Two-dimensional gel analysis of soluble proteins: charaterization of guinea pig exocrine pancreatic proteins. *J Biol Chem*. 1975;250:5375-5385.
19. Gibson G, Muse SV. *A Primer of Genome Science*. Sunderland, Mass: Sinauer Associates; 2002:190, 195.
20. Bjellqvist B, Ek K, Righetti PG, et al. Isoelectric focusing in immobilized pH gradients: principle, methodology and some applications. *J Biochem Biophys Methods*. 1982;6:317-339.
21. Gorg A, Postel W, Gunther S. The current state of two-dimensional electrophoresis with immobilized pH gradients. *Electrophoresis*. 1988;9:531-546.
22. Rabilloud T, Adessi C, Giraudel A, Lunardi J. Improvement of the solubilization of proteins in two-dimensional electrophoresis with immobilized pH gradients. *Electrophoresis*. 1997;18:307-316.
23. Garrels JI, McLaughlin CS, Warner JR, et al. Proteome studies of *Saccharomyces cerevisiae*: identification and characterization of abundant proteins. *Electrophoresis*. 1997;18:1347-1360.
24. Gorg A, Obermaier C, Boguth G, et al. The current state of two-dimensional electrophoresis with immobilized pH gradients. *Electrophoresis*. 2000;21:1037-1053.
25. Wildgruber R, Harder A, Obermaier C, et al. Towards higher resolution: two-dimensional electrophoresis of *Saccharomyces cerevisiae* proteins using overlapping narrow immobilized pH gradients. *Electrophoresis*. 2000;21:2610-2616.
26. Fey SJ, Larsen PM. 2D or not 2D: two-dimensional gel electrophoresis. *Curr Opin Chem Biol*. 2001;5:26-33.
27. Berggren K, Chernokalskaya E, Steinberg TH, et al. Background-free, high sensitivity staining of proteins in one- and two-dimensional sodium dodecyl sulfate-polyacrylamide gels using a luminescent ruthenium complex. *Electrophoresis*. 2000;21:2509-2521.
28. Rabilloud T. Mechanisms of protein silver staining in polyacrylamide gels: a 10-year synthesis. *Electrophoresis*. 1990;11:785-794.
29. Lopez MF. Better approaches to finding the needle in a haystack: optimizing proteome analysis through automation. *Electrophoresis*. 2000;21:1082-1093.
30. Hanash SM. Biomedical applications of two-dimensional electrophoresis using immobilized pH gradients: current status. *Electrophoresis*. 2000;21:1202-1209.
31. Ong SE, Pandey A. An evaluation of the use of two-dimensional gel electrophoresis in proteomics. *Biomol Eng*. 2001;18:195-205.
32. Anderson NL, Anderson NG. Proteome and proteomics: new technologies, new concepts, and new words. *Electrophoresis*. 1998;19: 1853-1861.
33. Edman P, Begg G. A protein sequenator. *Eur J Biochem*. 1967;1: 80-91.
34. Arnott D. Basics of triple-stage quadrupole/ion-trap mass spectrometry: precursor, product and neutral-loss scanning. electrospray ionisation and nanospray ionisation. In: James P, ed. *Proteome Research: Mass Spectrometry*. Berlin, Germany: Springer-Verlag; 2001:11-31.

35. Beavis RC, Fenyö D. Database searching with mass-spectrometric information. In: Blackstock W, Mann M, eds. *Proteomics: A Trends Guide*. London, England: Elsevier; 2000:22-27.
36. Blackstock W. Trends in automation and mass spectrometry for proteomics. In: Blackstock W, Mann M, eds. *Proteomics: A Trends Guide*. London, England: Elsevier; 2000:12-17.
37. James P. Mass spectrometry and the proteome. In: James P, ed. *Proteome Research: Mass Spectrometry*. Berlin, Germany: Springer-Verlag; 2001:1-9.
38. Johnson RS. Automated interpretation of peptide tandem mass spectra and homology searching. In: James P, ed. *Proteome Research: Mass Spectrometry*. Berlin, Germany: Springer-Verlag; 2001:167-185.
39. Spengler B. The basics of matrix-assisted laser desorption, ionisation time-of-flight mass spectrometry and post-source decay analysis. In: James P, ed. *Proteome Research: Mass Spectrometry*. Berlin, Germany: Springer-Verlag; 2001:33-53.
40. Stahl DC, Lee TD. Data-controlled micro-scale liquid chromatography—tandem mass spectrometry of peptides and proteins: strategies for improved sensitivity, efficiency and effectiveness. In: James P, ed. *Proteome Research: Mass Spectrometry*. Berlin, Germany: Springer-Verlag; 2001:55-74.
41. Staudenmann W, James P. Interpreting peptide tandem mass-spectrometry fragmentation spectra. In: James P, ed. *Proteome Research: Mass Spectrometry*. Berlin, Germany: Springer-Verlag; 2001:143-166.
42. Tabb DL, Eng JK, Yates JR III. Protein identification by SEQUEST. In: James P, ed. *Proteome Research: Mass Spectrometry*. Berlin, Germany: Springer-Verlag; 2001:125-142.
43. Yates JR III. Mass spectrometry: from genomics to proteomics. *Trends Genet*. 2000;16:5-8.
44. Hillenkamp F, Karas M, Beavis RC, Chait BT. Matrix-assisted laser desorption/ionization mass spectrometry of biopolymers. *Anal Chem*. 1991;63:1193A-1203A.
45. Karas M, Hillenkamp F. Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Anal Chem*. 1988;60:2299-2301.
46. Banks JF Jr, Whitehouse CM. Electrospray ionization mass spectrometry. *Methods Enzymol*. 1996;270:486-519.
47. Fenn JB, Mann M, Meng CK, Wong SF, Whitehouse CM. Electrospray ionization for mass spectrometry of large biomolecules. *Science*. 1989;246:64-71.
48. Wilkins MR, Gooley AA. Protein identification in proteome projects. In: Wilkins MR, Williams KL, Appel RD, Hochstrasser DF, eds. *Proteome Research: New Frontiers in Functional Genomics*. Berlin, Germany: Springer-Verlag; 1997:35-64.
49. Henzel WJ, Billeci TM, Stults JT, Wong SC, Grimley C, Watanabe C. Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases. *Proc Natl Acad Sci U S A*. 1993;90:5011-5015.
50. James P, Quadroni M, Carafoli E, Gonnet G. Protein identification by mass profile fingerprinting. *Biochem Biophys Res Commun*. 1993;195:58-64.
51. Mann M, Hojrup P, Roepstorff P. Use of mass spectrometric molecular weight information to identify proteins in sequence databases. *Biol Mass Spectrom*. 1993;22:338-345.
52. Yates JR III, Speicher S, Griffin PR, Hunkapiller T. Peptide mass maps: a highly informative approach to protein identification. *Anal Biochem*. 1993;214:397-408.
53. Shevchenko A, Jensen ON, Podtelejnikov AV, et al. Linking genome and proteome by mass spectrometry: large-scale identification of yeast proteins from two dimensional gels. *Proc Natl Acad Sci U S A*. 1996;93:14440-14445.
54. Jensen ON, Podtelejnikov A, Mann M. Delayed extraction improves specificity in database searches by matrix-assisted laser desorption/ionization peptide maps. *Rapid Commun Mass Spectrom*. 1996;10:1371-1378.
55. Burlingame AL. Characterization of protein glycosylation by mass spectrometry. *Curr Opin Biotechnol*. 1996;7:4-10.
56. Roepstorff P. Mass spectrometry in protein studies from genome to function. *Curr Opin Biotechnol*. 1997;8:6-13.
57. Mann M, Wilm M. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal Chem*. 1994;66:4390-4399.
58. Yates JR III, Eng JK, McCormack AL. Mining genomes: correlating tandem mass spectra of modified and unmodified peptides to sequences in nucleotide databases. *Anal Chem*. 1995;67:3202-3210.
59. McCormack AL, Schieltz DM, Goode B, et al. Direct analysis and identification of proteins in mixtures by LC/MS/MS and database searching at the low-femtomole level. *Anal Chem*. 1997;69:767-776.
60. Hunt DF, Yates JR III, Shabanowitz J, Winston S, Hauer CR. Protein sequencing by tandem mass spectrometry. *Proc Natl Acad Sci U S A*. 1986;83:6233-6237.
61. Appel RD. Interfacing and integrating databases. In: Wilkins MR, Williams KL, Appel RD, Hochstrasser DF, eds. *Proteome Research: New Frontiers in Functional Genomics*. Berlin, Germany: Springer-Verlag; 1997:149-175.
62. Bairoch A. Proteome databases. In: Wilkins MR, Williams KL, Appel RD, Hochstrasser DF, eds. *Proteome Research: New Frontiers in Functional Genomics*. Berlin, Germany: Springer-Verlag; 1997:93-132.
63. Langen H, Berndt P. Proteomics databases. In: James P, ed. *Proteome Research: Mass Spectrometry*. Berlin, Germany: Springer-Verlag; 2001:229-257.
64. Bairoch A, Apweiler R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res*. 2000;28:45-48.
65. Gleich GJ, Leiferman KM, Pardanani A, Tefferi A, Butterfield JH. Treatment of hypereosinophilic syndrome with imatinib mesilate. *Lancet*. 2002;359:1577-1578.

Primer on Medical Genomics Part V will appear in the January issue.